#5 2 Jul 02

SSB Letter (2/01)

JC971 U.S. PTO
09/899536
07/06/01

Attorney Docket No. 1454.1074

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:

Martin HOLZAPFEL

Application No.: Group Art Unit:

Filed: (concurrently) Examiner:

For: METHOD FOR GENERATING A STATISTIC FOR PHONE LENGTHS AND METHOD FOR DETERMINING THE LENGTH OF INDIVIDUAL PHONES FOR SPEECH SYNTHESIS

SUBMISSION OF CERTIFIED COPY OF PRIOR FOREIGN APPLICATION IN ACCORDANCE WITH THE REQUIREMENTS OF 37 C.F.R. §1.55

Assistant Commissioner for Patents
Washington, D.C. 2023l

Sir:

In accordance with the provisions of 37 C.F.R. §1.55, the applicant(s) submit(s) herewith a certified copy of the following foreign application:

German Patent Application No. 10033104.1

Filed: 7 July 2000

It is respectfully requested that the applicant(s) be given the benefit of the foreign filing date(s) as evidenced by the certified papers attached hereto, in accordance with the requirements of 35 U.S.C. §119.

Respectfully submitted,

STAAS & HALSEY LLP

Date: 7/6/01

By: Richard A. Gollhofer
Richard A. Gollhofer
Registration No. 31,106

700 11th Street, N.W., Ste. 500
Washington, D.C. 20001

BUNDESREPUBLIK DEUTSCHLAND

Prioritätsbescheinigung über die Einreichung einer Patentanmeldung

Aktenzeichen:	100 33 104.1
Anmeldetag:	07. Juli 2000
Anmelder/Inhaber:	Siemens Aktiengesellschaft, München/DE
Bezeichnung:	Verfahren zum Erzeugen einer Statistik von Phondauern und Verfahren zum Ermitteln der Dauer einzelner Phone für die Sprachsynthese
IPC:	G 10 L 13/00

Die angehefteten Stücke sind eine richtige und genaue Wiedergabe der ursprünglichen Unterlagen dieser Patentanmeldung.

München, den 12. April 2001
Deutsches Patent- und Markenamt
Der Präsident
Im Auftrag

Agurks

Beschreibung

Verfahren zum Erzeugen einer Statistik von Phondauern und Verfahren zum Ermitteln der Dauer einzelner Phone für die Sprachsynthese

Die vorliegende Erfindung betrifft ein Verfahren zum Erzeugen einer Statistik von Phondauern und ein Verfahren zum Ermitteln der Dauer einzelner Phone für die Sprachsynthese.

Im Sinne der vorliegenden Anmeldung ist ein Phonem die kleinste bedeutungsunterscheidende, aber nicht selbstbedeutungstragende sprachliche Einheit (z.B. b in Bein im Unterschied zu p in Pein). Ein Phon ist hingegen der ausgesprochene Laut eines Phonems.

Verfahren zum Erzeugen einer Statistik von Phondauern, wobei auf Grundlage dieser Statistik bei der synthetischen Spracherzeugung die Phondauern gesteuert werden können, sind bekannt. Bei derartigen Verfahren wird ein von einem Sprecher gesprochener Text aufgezeichnet und der aufgezeichnete Text in einzelne Phone segmentiert. Von den einzelnen Phonen wird die Lautlänge bestimmt. Diese Phondauern wird in einer Statistik erfasst, wobei die Statistik eine Liste von Triphonen aufweist. Ein Triphon ist ein Kluster von einem oder mehreren Phonemen mit dem jeweiligen rechten und linken Kontext.

Bei den bekannten Verfahren wird jeweils einem Phonem der Triphone in ihrem links-rechts Kontext eine mittlere Phonlänge bzw. Lautdauer zugeordnet. Diese Phondauern wird aus allen Phonen des gesprochenen Textes ermittelt, die im gleichen Kontext im gesprochenen Text wie in dem jeweiligen Triphon vorkommen, das heißt deren benachbarte Phone korrespondieren zu den benachbarten Phonemen im Triphon.

Bei den bekannten Verfahren zum Ermitteln der Dauer einzelner Phone für die Sprachsynthese werden den Phonemen des zu syn-

thetisierenden Textes die jeweils mittlere Lautdauer des Phonems der Statistik zugeordnet, dessen Kontext im Triphon dem Kontext des Phonems im zu synthetisierenden Textes entspricht. Ist z.B. die Phondauern des Phonems „b" des Wortes „aber" zu Ermitteln, so wird bei dem bekannten Verfahren dem Phonem „b" diejenige Phondauern zugeordnet, die in der Statistik dem Phonem „b" im Triphon „abe" zugeordnet ist. Die Kontexte des Triphons und im zu synthetisierenden Text sind hier jeweils identisch.

Der Erfindung liegt die Aufgabe zugrunde, ein Verfahren zum Erzeugen einer Statistik von Phondauern, wobei auf Grundlage dieser Statistik bei der synthetischen Spracherzeugung die Phondauern gesteuert werden können, und ein Verfahren zum Ermitteln der Dauer einzelner Phone für die Sprachsynthese zu schaffen, wodurch eine Sprachsynthese mit natürlicherer Aussprache als bei bekannten Verfahren erzielt werden soll.

Die Aufgabe wird mit einem Verfahren zum Erzeugen einer Statistik von Phondauern mit den Merkmalen des Anspruchs 1 und durch ein Verfahren zum Ermitteln der Dauer einzelner Phone mit den Merkmalen des Anspruchs 11 gelöst. Vorteilhafte Ausgestaltungen der Erfindung sind in den Unteransprüchen angegeben.

Das erfindungsgemäße Verfahren zum Erzeugen einer Statistik von Phondauern auf Grundlage der bei der synthetischen Spracherzeugung die Phondauern gesteuert werden können, umfasst folgende Schritte:

- Zuordnen von Phonen eines in Phone segmentierten gesprochenen und aufgezeichneten Textes zu Phonemen von vorbestimmten Primärklustern, die aus mehreren Phonemen zusammengesetzt sind, wobei jeweils ein Phon einem Phonem eines Primärklusters zugeordnet wird, wenn es im gesprochenen Text zu einem im Kontext des Phonems des Primärklusters identischen oder ähnlichen Kontext auftritt,

- Erstellen einer Primärstatistik, die zumindest die mittlere Phondauern aller Phone, die dem jeweiligen Phonem eines Primärklusters zugeordnet sind, umfasst,

- Zuordnen von Phonen des gesprochenen und aufgezeichneten Textes zu Phonemen zu vorbestimmten Sekundärklustern, die aus Phonemen zusammengesetzt sind, wobei zumindest die Anzahl Phoneme einiger Sekundärkluster sich von der Anzahl der Phoneme der Primärkluster unterscheidet, wobei jeweils ein Phon einem Phonem eines Sekundärklusters zugeordnet wird, wenn es im gesprochenen Text zu einem im Kontext des Phonems des Sekundärklusters identischen Kontext auftritt,

- Erstellen einer Sekundärstatistik, die zumindest die mittlere Phondauern aller Phone, die dem jeweiligen Phonem eines Sekundärklusters zugeordnet sind, umfasst.

Die durch das erfindungsgemäße Verfahren erzeugte Statistik besteht somit aus einer Primärstatistik und einer Sekundärstatistik. Die Primärstatistik kann auf Primärkluster mit z.B. jeweils drei Phonemen beruhen, so dass sie der eingangs erläuterten Statistik auf Basis von Triphonen entspricht. Die Sekundärstatistik ist eine weitere Statistik auf Basis von Sekundärklustern, die sich in der Anzahl der Phoneme zumindest teilweise von der Anzahl der Phoneme der Primärkluster unterscheiden. Hierdurch wird eine sprachspezifischere Statistik zur Phondauern erzielt.

So können z.B. die Primärkluster drei Phoneme und die Sekundärkluster vier Phoneme umfassen, wodurch ein größerer Kontext (vier Phoneme gegenüber drei Phonemen) bei der Ermittlung der mittleren Phondauern berücksichtigt wird, so dass durch eine wesentlich sprachspezifischere Auswertung erzielt wird.

Nach einer bevorzugten Ausführungsform der Erfindung besitzen die Primärkluster eine konstante Anzahl Phoneme, wohingegen

die Anzahl der Phoneme der Sekundärkluster variabel ist. So können z.B. die Primärkluster jeweils drei Phoneme und die Sekundärkluster jeweils alle Phoneme eines Wortes umfassen. Mit Hilfe dieser Sekundärkluster wird dann eine wortspezifische Auswertung der Phondauern erzielt, die wesentlich präziser ist, als die auf Grundlage der Triphone.

Nach einer bevorzugten Ausführungsform der Erfindung werden in der Sekundärstatistik nur Sekundärkluster erfasst, deren Häufigkeit im Text größer oder gleich einer vorbestimmten Mindesthäufigkeit ist. Hierdurch wird sichergestellt, dass in der Statistik nicht signifikante Häufigkeiten nicht berücksichtigt werden. So ist es zweckmäßig, Wörter, die in dem Text, auf dem die Statistik beruht, lediglich einmal oder zweimal vorkommen, nicht zu berücksichtigen.

Das erfindungsgemäße Verfahren zum Ermitteln der Dauer einzelner Phone für die Sprachsynthese beruht auf einer derartigen eine Primärstatistik und eine Sekundärstatistik umfassenden Statistik von Phondauern. Dieses Verfahren umfasst folgende Schritte:

- Bestimmen, ob das in Sprache umzusetzende Phonem, für das die Phondauern zu ermitteln ist, Bestandteil eines Sekundärklusters ist,

- Zuordnen der mittleren Phondauern (d), die in der Sekundärstatistik dem entsprechenden Phonem in dem jeweiligen Sekundärkluster zugeordnet ist, falls das Phonem Bestandteil eines Sekundärklusters ist, und

- Zuordnen der mittleren Phondauern (d), die in der Primärstatistik dem entsprechenden Phonem in dem jeweiligen Primärkluster zugeordnet ist, falls das Phonem nicht Bestandteil eines Sekundärklusters ist.

Bei diesem Verfahren wird bevorzugt die sprachspezifischere Sekundärstatistik bei der Ermittlung der Phondauern ausgewertet. Hierbei ist zu berücksichtigen, dass beim Erzeugen der Sekundärstatistik lediglich identische Kontexte zwischen dem Sekundärkluster und dem entsprechenden Abschnitt in dem gesprochenen und aufgezeichneten Text, auf dem die Statistiken beruhen, berücksichtigt werden, wohingegen bei der Primärstatistik auch ähnliche Kluster zu berücksichtigen sind, falls keine identische Übereinstimmung vorhanden ist. Dies ist ein weiterer Grund, weshalb zunächst versucht wird, die Sekundärstatistik auszuwerten, bevor auf die Primärstatistik zurückgegriffen wird.

Gemäß einer bevorzugten Weiterbildung des Verfahrens zum Ermitteln der Dauer einzelner Phone wird die Standardabweichung der einzelnen mittleren Phondauern berücksichtigt. Dies bewirkt eine weitere Anpassung an eine natürliche Aussprache.

Die Erfindung wird nachfolgend beispielhaft anhand der beiliegenden Zeichnungen näher erläutert. In denen zeigen schematisch:

Fig. 1 einen allgemeinen Überblick über die Abläufe bei der Erzeugung einer Statistik von Phondauern in einem Flussdiagramm,

Fig. 2 die Verfahrensschritte zur statistischen Auswertung einer Sprachaufzeichnung zur Erzeugung einer Statistik von Phondauern,

Fig. 3 ein Verfahren zum Ermitteln der Dauer einzelner Phone für die Sprachsynthese in einem Flussdiagramm, und

Fig. 4 ein Computersystem zum Ausführen der erfindungsgemäßen Verfahren in einem Blockschaltbild.

Fig. 1 zeigt die grundlegenden Abläufe für ein Verfahren zum Erzeugen einer Statistik von Phondauern, auf deren Grundlage bei der synthetischen Spracherzeugung die Phondauern gesteuert werden kann.

Das Verfahren beginnt mit dem Schritt S1 und im Schritt S2 wird ein vorbestimmter Trainingstext von einem Sprecher gesprochen und aufgezeichnet. Die Aufzeichnung erfolgt mittels eines Mikrofons, das die akustischen Sprachsignale in korrespondierende elektrische Sprachsignale wandelt.

Das aufgezeichnete Sprachsignal wird im Schritt S3 in einzelne Phone segmentiert. Das Segmentieren des Sprachsignals in die einzelnen Phone wird oftmals von einem Sprachexperten manuell durchgeführt. Es sind auch voll- und teilautomatische Verfahren bekannt, die in der Regel auf einem HMM (Hidden-Markow-Model) Algorithmus beruhen.

Im Schritt S4 werden die einzelnen Phone statistisch ausgewertet, wobei deren Dauer bestimmt wird. Phondauern von Phonen, die dem gleichen Phonem im gleichen oder ähnlichen Kontext zugeordnet sind, werden statistisch ausgewertet, indem deren Mittelwerte und Standardabweichungen berechnet werden.

Im Schritt S5 wird dieses Verfahren beendet.

Die erfindungsgemäß auszuführenden Verfahrensschritte bei der statistischen Auswertung (S4) sind in Fig. 2 in einem Flussdiagramm dargestellt. Mit dem Schritt S6 beginnt das statistische Auswerteverfahren. Zunächst werden die einzelnen Phone des Trainingstextes einem Primärkluster zugeordnet. Im vorliegenden Ausführungsbeispiel ist das Primärkluster ein aus drei Phonemen bestehendes Triphon. Ein Phon des Trainingstextes wird demjenigen Triphon zugeordnet, dessen mittleres Phonem dem Phon des Trainingstextes entspricht und das den gleichen Kontext wie der Abschnitt des Trainingstextes in dem das zuzuordnende Phon angeordnet ist, aufweist. Dies bedeutet,

dass die zum mittleren Phonem des Triphons benachbarten Phoneme den benachbarten Phonen des zuzuordnenden Phones des Trainingstextes entsprechen. Soll z.B. das Phon des Phonems „f" des Wortes „Anfang" einem solchen Primärkluster zugeordnet werden, so wird dieses Phon dem Phonem „f" im Triphon „nfa" zugeordnet, da die beiden benachbarten Phoneme „n" (links) und „a" (rechts) den entsprechenden Phonen von „n" und „a" im Trainingstext entsprechen.

Die Primärkluster sind in einer vorab festgelegten Liste gespeichert. Sind die Primärkluster Triphone, so umfasst eine solche Liste typischerweise 1500 bis 2000 Triphone. In dieser Liste sind die am häufigsten auftretenden Permutationen von drei aufeinanderfolgenden Phonemen enthalten. Selten und ähnlich klingende Permutationen werden in einem Kluster zusammengefasst. So können z.B. die Triphone „ter" und „der" in einem Kluster zusammengefasst sein.

Bei der Zuordnung nach dem Schritt S7 werden somit die Phone den jeweiligen Phonemen im gleichen oder ähnlichen Kontext zugeordnet.

Am Ende dieses Zuordnungsvorganges sind der Liste der Primärkluster alle Phone des Trainingstextes zugeordnet, das heißt, dass eine Liste vorliegt, in der zu jedem Primärkluster die entsprechenden Phone des Trainingstextes gespeichert sind.

Im Schritt S8 wird die mittlere Phondauern d´ und die Standardabweichung G für das jeweils mittlere Phonem eines jedem aus drei Phonemen bestehenden Primärklusters berechnet. Hierbei werden die Lautdauern der einzelnen einem Primärkluster zugeordneten Phone gemittelt und als mittlere Lautdauer gespeichert und die entsprechende Standardabweichung G berechnet.

Mit dem Schritt S8 wird somit eine Primärstatistik erzeugt, die im wesentlicher der eingangs erörterten, aus dem Stand der Technik bekannten Statistik entspricht.

Im Schritt S9 werden die einzelnen Phone Sekundärklustern zugeordnet. Im vorliegenden Ausführungsbeispiel umfassen die Sekundärkluster jeweils alle Phoneme eines Wortes. Die Länge der Sekundärkluster ist somit variabel. Bei der Zuordnung der Phone zu den Sekundärklustern werden die Wörter des Trainingstextes ermittelt und die einzelnen Phone dieser Wörter werden den korrespondierenden Phonemen der entsprechenden Sekundärkluster zugeordnet. Ein wesentlicher Unterschied gegenüber dem Schritt S7 ist, dass hier nicht nur ein Phon einem Kluster zugeordnet wird, sondern alle Phone eines Wortes werden den entsprechenden Phonemen des Sekundärkluster zugeordnet, das heißt, dass allen Phonemen des Sekundärklusters jeweils ein Phon zugeordnet wird. Im Schritt S10 wird geprüft, ob den Phonemen der Sekundärkluster jeweils mindestens drei Phone des Trainingstextes zugeordnet worden sind. Ist dies nicht der Fall, bedeutet dies, dass das entsprechende Wort im Trainingstext weniger als dreimal vorkommt und deshalb nicht statistisch signifikant ist. Sekundärkluster, denen weniger als drei Wörter des Trainingstextes zugeordnet worden sind, werden gelöscht.

Im vorliegenden Ausführungsbeispiel beträgt die geforderte Häufigkeit für die Signifikanz drei. Zur Erzielung einer größeren statistischen Sicherheit kann es zweckmäßig sein, einen entsprechend höheren Wert anzusetzen.

Im Schritt S11 wird die mittlere Phondauern d´ und die Standardabweichung G für ein jedes Phonem des Sekundärklusters berechnet und abgespeichert. Als Ergebnis des Schrittes S11 wird eine Sekundärstatistik auf Grundlage der Sekundärkluster erhalten.

Im Schritt S12 wird das Auswerteverfahren beendet.

Mit dem in Fig. 2 gezeigten Ausführungsbeispiel wird eine Statistik erhalten, die wesentlich sprachspezifischer ist, da die einzelnen Phondauern sehr stark von dem entsprechenden Kontext abhängen und ein wesentlich präziserer Kontext durch den Kontext eines gesamten Wortes berücksichtigt wird, falls dies statistisch möglich ist. Wird auf Grundlage einer solchen zweistufigen Statistik die Lautdauer für eine Sprachsynthese bestimmt, so ermöglicht dies eine wesentlich natürlichere Synthese der Sprache.

Im Rahmen der Erfindung können sowohl andere Primärkluster und Sekundärkluster verwendet werden. Insbesondere ist es z.B. möglich Sekundärkluster mit einer konstanten Länge von z.B. vier Phonemen zu verwenden. Es könnte jedoch auch zweckmäßig sein, bei bestimmten Anwendungen, wesentlich längere Sekundärkluster zu verwenden, die z.B. eine vollständige Phrase, einen vollständigen Satz oder einen ganzen Absatz umfassen können. Je länger die Sekundärkluster gewählt werden, desto spezieller sollte das Anwendungsgebiet der Sprachsynthese sein. Ein typisches Beispiel für ein sehr spezielles Anwendungsgebiet einer Sprachsynthese ist ein Navigationssystem für Kraftfahrzeuge, bei dem wiederholt sehr ähnliche Sätze und Satzstrukturen erzeugt werden.

In Fig. 3 ist ein Verfahren zum Ermitteln einzelner Phone für die Sprachsynthese schematisch in einem Flussdiagramm dargestellt.

Ausgangspunkt des Verfahrens ist, dass ein Phonem eines zu synthetisierenden Textes in ein Phon umgesetzt wird und die Dauer dieses Phons zu bestimmen ist.

Das Verfahren beginnt mit dem Schritt S13. Im Schritt S14 wird der Kontext des Phonems im Ausgangstext bestimmt. Hierbei wird zweckmäßigerweise der Umfang des Kontextes so gewählt, dass er der Länge des Sekundärklusters entspricht. Im

vorliegenden Ausführungsbeispiel wird der Kontext im Umfang eines Wortes bestimmt.

Im Schritt S15 wird geprüft, ob der im Schritt S14 ermittelte Kontext als Sekundärkluster in der Sekundärstatistik gespeichert ist. Ist dies der Fall, geht der Programmablauf auf den Schritt S16 über, mit dem die mittlere Phondauern d´ die dem Phonem des Sekundärklusters zugeordnet ist, der dem Phonem des Ausgangstextes entspricht, und die Phondauern und die Standardabweichung ausgelesen werden. Der Programmablauf geht dann auf den Schritt S17 über, bei dem die tatsächlich anzuwendende Phondauern d aus der mittleren Phondauern d´ und der Standardabweichung G gemäß folgender Formel berechnet wird:

$$d = d' + G \cdot s,$$

wobei s ein Geschwindigkeitsskalierungsfaktor ist, der gemäß folgender Formel berechnet wird:

$$s = R_{rel} - 1,$$

wobei R_{rel} das Verhältnis der zu sprechenden Sprechgeschwindigkeit gegenüber der Sprechgeschwindigkeit ist, mit der der Text auf dem die Statistik beruht, gesprochen worden ist. Durch die Berücksichtigung der Standardabweichung werden Phone, die der Sprecher des Trainingstextes mit stark unterschiedlichen Längen ausgesprochen hat, entsprechend stark bei der Sprachsynthese variiert. Z.B. werden Plosiv-Laute, wie z.B. „k" sehr wenig variiert, weshalb sie eine sehr kleine Standardabweichung besitzen. Sie werden bei der Sprachsynthese entsprechend wenig variiert. Vokale, wie z.B. „a" werden stark variiert, weshalb sie eine entsprechend große Standardabweichung besitzen. Bei obigen Formeln ist zu berücksichtigen, dass der Geschwindigkeitsskalierungsfaktor s auch negative Werte annehmen kann, wodurch die Phondauern gegenüber der mittleren Phondauern entsprechend verkürzt wird.

Ergibt die Abfrage im Schritt S15 hingegen, dass der im Schritt S14 ermittelte Kontext nicht in der Sekundärstatistik enthalten ist, so geht der Verfahrensablauf auf den Schritt S18 über. Im Schritt S18 wird geprüft, ob der Abschnitt des Kontextes im Bereich des umzusetzenden Phonems identisch zu einem Primärkluster der Primärstatistik ist. Ist dies der Fall, geht der Verfahrensablauf auf den Schritt S19 über. Im Schritt S19 wird die mittlere Phondauern und die Standardabweichung des mittleren Phonems des entsprechenden Primärklusters ausgelesen. Der Verfahrensablauf geht dann auf den Schritt S17 über, mit dem in der oben erläuterten Weise die tatsächlich anzuwendende Phondauern berechnet wird.

Ergibt die Abfrage im Schritt S18, dass zu dem Kontext des Ausgangstextes kein identisches Primärkluster in der Primärstatistik vorhanden ist, so geht der Verfahrensablauf auf den Schritt S20 über, in dem ein Primärkluster bestimmt wird, das dem Kontext klanglich möglichst ähnlich ist.

Im darauffolgenden Schritt S21 werden die mittlere Phondauern und die Standardabweichung des mittleren Phonems dieses Primärklusters ausgelesen. Der Verfahrensablauf geht dann auf den Schritt S17 über.

Nach Ausführung des Schrittes S17 wird das Verfahren zum Ermitteln der Dauer eines Phons eines Phonems eines Ausgangstextes im Schritt S18 beendet.

Das erfindungsgemäße Verfahren zum Bestimmen der Phondauern für die Sprachsynthese ist somit ein zweistufiges Verfahren, bei dem zunächst versucht wird, mittels der Sekundärstatistik eine mittlere Phondauern zu ermitteln, die auf einem speziellen Kontext (hier: Wortlänge) beruht, wodurch eine Lautdauer ermittelt wird die der natürlichen Sprechweise wesentlich ähnlicher ist, als die auf Grund der Primärstatistik ermittelte Phondauern. Sollte diese Phondauernbestimmung mittels der Sekundärstatistik nicht möglich sein, so wird auf die

Primärstatistik zurückgegriffen, die grundsätzlich immer anwendbar ist.

Insbesondere die Kombination des Verfahrens zum Erzeugen der Statistik und des Verfahrens zum Ermitteln der Phondauern stellt ein im wesentlichen rein statistisches Verfahren zur Ermittlung der Phondauern dar, das im wesentlichen ohne Expertenwissen erstellt und angewendet werden kann. Bei dem oben beschriebenen Ausführungsbeispiel wird z.B. lediglich bei der Segmentierung der Sprachaufzeichnung Expertenwissen eingesetzt, wobei dieser Schritt mittels bekannter Verfahren auch automatisierbar ist.

Die erfindungsgemäßen Verfahren sind so einfach zu implementieren und zu trainieren. Dennoch haben erste Versuche mit Prototypen gezeigt, dass sie bei der Sprachsynthese eine wesentliche Steigerung der Sprachqualität bewirken, da die Phondauern durch das Vorsehen der Sekundärstatistik sprachspezifischer ermittelt wird.

Die oben beschriebenen Verfahren können als Computerprogramme realisiert werden, die selbständig auf einem Computer zum Erzeugen der Statistik bzw. zum Ermitteln der Phondauern ablaufen. Sie stellen somit automatisch ausführbare Verfahren dar.

Die Computerprogramme können auch auf elektrisch lesbaren Datenträgern gespeichert werden und so auf andere Computersysteme übertragen werden.

Ein zur Anwendung des erfindungsgemäßen Verfahrens geeignetes Computersystem ist in Fig. 4 gezeigt. Das Computersystem 1 weist einen internen Bus 2 auf, der mit einem Speicherbereich 3, einer zentralen Prozessoreinheit 4 und einem Interface 5 verbunden ist. Das Interface 5 stellt über eine Datenleitung 6 eine Datenverbindung zu weiteren Computersystemen her. An dem internen Bus 2 sind ferner eine akustische Ausgabeeinheit 7, eine grafische Ausgabeeinheit 8 und eine Eingabeeinheit 9

angeschlossen. Die akustische Ausgabeeinheit 7 ist mit einem Lautsprecher 10, die grafische Ausgabeeinheit 8 mit einem Bildschirm 11 und die Eingabeeinheit 9 mit einer Tastatur 12 verbunden. An dem Computersystem 1 können über die Datenleitung 6 und das Interface 5 Sprachaufzeichnungen eines Textes übertragen werden, die im Speicherbereich 3 abgespeichert werden. Der Speicherbereich 3 ist in mehrere Bereiche unterteilt, in denen Sprachaufzeichnungen, Audiodateien, Anwendungsprogramme zum Durchführen der erfindungsgemäßen Verfahren und weitere Anwendungs- und Hilfsprogramme gespeichert sind. Die Sprachdateien werden mit vorbestimmten Programmpaketen analysiert und in die einzelnen Phone segmentiert. Danach wird das erfindungsgemäße Verfahren zum Erzeugen einer Statistik ausgeführt, wobei als Ergebnis die Primär- und Sekundärstatistik vorliegen.

Ein beispielsweise über die Datenleitung 6 und das Interface 5 im Speicherbereich 3 abgespeicherter Text kann dann in eine Audiodatei umgesetzt werden, wobei die Phondauern mittels des erfindungsgemäßen Verfahrens (Fig. 3) auf Grundlage der Primär- und Sekundärstatistik bestimmt werden.

Eine so erzeugte Audiodatei wird über den internen Bus 2 zur akustischen Ausgabeeinheit 7 übertragen und von dieser am Lautsprecher 10 als Sprache ausgegeben.

Patentansprüche

1. Verfahren zum Erzeugen einer Statistik von Phondauern, wobei auf Grundlage dieser Statistik bei der synthetischen Spracherzeugung die Phondauern gesteuert werden können, umfassend folgende Schritte:

- Zuordnen von Phonen eines in Phone segmentierten gesprochenen und aufgezeichneten Textes zu Phonemen von vorbestimmten Primärklustern, die aus mehreren Phonemen zusammengesetzt sind, wobei jeweils ein Phon einem Phonem eines Primärklusters zugeordnet wird, wenn es im gesprochenem Text zu einem dem Kontext des Phonems des Primärklusters identischen oder ähnlichen Kontext auftritt,
- Erstellen einer Primärstatistik, die zumindest die mittlere Phondauern aller Phone, die dem jeweiligen Phonem eines Primärklusters zugeordnet sind, umfasst,
  g e k e n n z e i c h n e t d u r c h
- Zuordnen von Phonen des gesprochenen und aufgezeichneten Textes zu Phonemen von vorbestimmten Sekundärklustern, die aus Phonemen zusammengesetzt sind, wobei zumindest die Anzahl Phoneme einiger Sekundärkluster sich von der Anzahl der Phoneme der Primärkluster unterscheidet, wobei jeweils ein Phon einem Phonem eines Sekundärklusters zugeordnet wird, wenn es im gesprochenem Text zu einem dem Kontext des Phonems des Sekundärklusters identischen Kontext auftritt,
- Erstellen einer Sekundärstatistik, die zumindest die mittlere Phondauern aller Phone, die dem jeweiligen Phonem eines Sekundärklusters zugeordnet sind, umfasst.

2. Verfahren zum Erzeugen einer Statistik von Phondauern nach Anspruch 1,
d a d u r c h g e k e n n z e i c h n e t,
dass die Anzahl der Phoneme der Primärkluster konstant ist und die Anzahl z.B. gleich 3 ist.

3. Verfahren zum Erzeugen einer Statistik nach Anspruch 1 oder 2,

d a d u r c h g e k e n n z e i c h n e t,
dass die Anzahl der Phoneme des Sekundärklusters variabel ist und die Sekundärkluster z.B. jeweils die Phoneme eines Wortes umfassen.

4. Verfahren zum Erzeugen einer Statistik nach einem der Ansprüche 1 bis 3,
d a d u r c h g e k e n n z e i c h n e t,
dass die Primärstatistik und die Sekundärstatistik jeweils die Standardabweichung der jeweiligen Phondauern umfassen.

5. Verfahren zum Erzeugen einer Statistik nach einem der Ansprüche 1 bis 4,
d a d u r c h g e k e n n z e i c h n e t,
dass mit der Sekundärstatistik nur Sekundärkluster erfasst werden, deren Häufigkeit im Text größer oder gleich einer vorbestimmten Mindesthäufigkeit ist.

6. Verfahren zum Erzeugen einer Statistik nach einem der Ansprüche 1 bis 5,
d a d u r c h g e k e n n z e i c h n e t,
dass die Mindesthäufigkeit zumindest 3 beträgt und vorzugsweise im Bereich von 3 bis 10 liegt.

7. Verfahren zum Erzeugen einer Statistik nach einem der Ansprüche 1 bis 6,
d a d u r c h g e k e n n z e i c h n e t,
dass die Zuordnung der Phone zu Phonemen der Primärkluster mittels einer vorbestimmten Liste von in Primärklustern gruppierten Phonemen erfolgt, wobei die Phone den einzelnen Phonemen der Primärkluster der Liste zugeordnet werden und die einzelnen Zuordnungen abgespeichert werden.

8. Verfahren nach Anspruch 7,
d a d u r c h g e k e n n z e i c h n e t,
dass zu den einzelnen Phonemen der Primärklustern der Liste auf Grundlage der abgespeicherten Zuordnungen jeweils die

mittlere Phondauern (d) und die Standardabweichung (G) der mittleren Phondauern berechnet werden.

9. Verfahren nach einem der Ansprüche 1 bis 8,
d a d u r c h g e k e n n z e i c h n e t,
dass die Zuordnung der Phone zu den Phonemen der Sekundärkluster mittels einer vorbestimmten Liste von in Sekundärklustern gruppierten Phonemen erfolgt, wobei die Phone den einzelnen Phonemen der Sekundärkluster der Liste zugeordnet werden und die einzelnen Zuordnungen abgespeichert werden.

10. Verfahren nach Anspruch 9,
d a d u r c h g e k e n n z e i c h n e t,
dass zu den einzelnen Phonemen der Sekundärkluster der Liste auf Grundlage der abgespeicherten Zuordnungen jeweils die mittlere Phondauern (d) und die Standardabweichung (G) der mittleren Phondauern berechnet werden.

11. Verfahren zum Ermitteln der Dauer einzelne Phone für die Sprachsynthese, mittels einer Statistik von Phondauern, die eine Primärstatistik und eine Sekundärstatistik aufweist, wobei die Primärstatistik in Primärkluster gruppierte Phoneme umfasst, und den einzelnen Phonemen der Primärkluster zumindest eine mittlere Phondauern zugeordnet ist, und
die Sekundärstatistik in Sekundärkluster gruppierte Phoneme umfasst, und den einzelnen Phonemen der Sekundärkluster zumindest eine mittlere Phondauern zugeordnet ist,
umfassend folgende Schritte:
- Bestimmen, ob das in Sprache umzusetzende Phonem, für das die Phondauern zu ermitteln ist, Bestandteil eines Sekundärklusters ist,
- Zuordnen der mittleren Phondauern (d), die in der Sekundärstatistik dem entsprechendem Phonem in dem jeweiligen Sekundärkluster zugeordnet ist, falls das Phonem Bestandteil eines Sekundärklusters ist, und
- Zuordnen der mittleren Phondauern (d), die in der Primärstatistik dem entsprechendem Phonem in dem jeweiligen Pri-

märkluster zugeordnet ist, falls das Phonem nicht Bestandteil eines Sekundärklusters ist.

12. Verfahren zum Ermitteln der Dauer der einzelnen Phone bei der Sprachsynthese mittels einer Statistik mit einem Verfahren nach einem der Ansprüche 1 bis 10 erzeugten Statistik.

13. Verfahren nach Anspruch 11 oder 12,
dadurch gekennzeichnet,
dass bei der Ermittlung der Dauer (d) der einzelnen Phone die Standardabweichungen (G) der in der Statistik gespeicherten mittleren Phondauern (d´) gemäß folgender Formel berücksichtigt werden

$$d = d' + G \cdot s,$$

wobei s ein Geschwindigkeitsskalierungsfaktor ist, der gemäß folgender Formel berechnet wird

$$s = R_{rel} - 1,$$

wobei R_{rel} das Verhältnis der zu sprechenden Sprechgeschwindigkeit gegenüber der Sprechgeschwindigkeit, mit der der Text auf dem die Statistik beruht, gesprochen worden ist.

14. Vorrichtung zum Erzeugen einer Statistik von Phondauern auf Grundlage der bei der synthetischen Spracherzeugung die Phondauern gesteuert werden können, mit

einem Computersystem (1), das einen Speicherbereich (3) aufweist, in dem ein Programm zum Ausführen eines Verfahrens nach einem der Ansprüche 1 bis 10 gespeichert ist.

15. Vorrichtung zum Ermitteln der Dauer einzelner Phone für die Sprachsynthese mit

einem Computersystem (1), das einen Speicherbereich (3) aufweist, in dem ein Programm zum Ausführen eines Verfahrens nach einem der Ansprüche 11 bis 13 gespeichert ist.

Zusammenfassung

Verfahren zum Erzeugen einer Statistik von Phondauern und Verfahren zum Ermitteln der Dauer einzelner Phone für die Sprachsynthese

Die vorliegende Erfindung betrifft ein Verfahren zum Erzeugen einer Statistik von Phondauern und ein Verfahren zum Ermitteln der Dauer einzelner Phone für die Sprachsynthese.

Erfindungsgemäß wird eine Primärstatistik vorgesehen, die beispielsweise auf Primärklustern (z.B. Triphonen) beruht und eine Sekundärstatistik, die auf Sekundärklustern (z.B. Phoneme von ganzen Wörtern) beruht. Beide Statistiken beinhalten mittlere Phondauern und beispielsweise die Standardabweichung der mittleren Phondauern. Bei der Ermittlung der Phondauern wird zunächst versucht, diese anhand der Sekundärstatistik, die sprachspezifischer ist, zu ermitteln. Falls dies nicht der Fall ist, wird auf die Primärstatistik zurückgegriffen, die immer anwendbar ist. Durch dieses zweistufige Verfahren wird eine Phondauer ermittelt, die einer natürlichen Sprache wesentlich besser entspricht, als dies mit dem bekannten einstufigen Verfahren möglich war.

(Figur 2)

2000E01067
2000P13225
SRZ

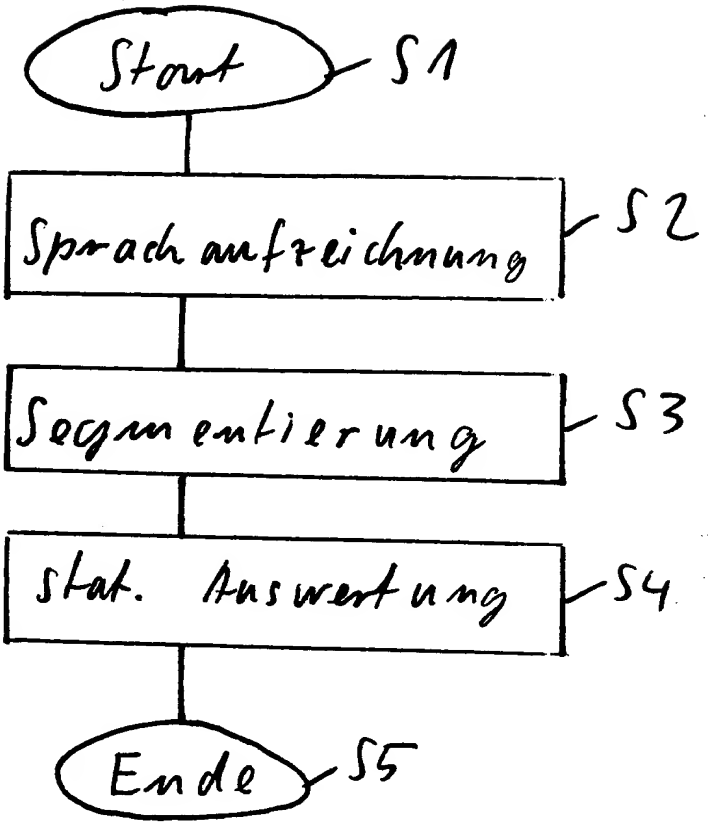

Fig. 1

2000 E 01067
2000 P 13225
SRZ

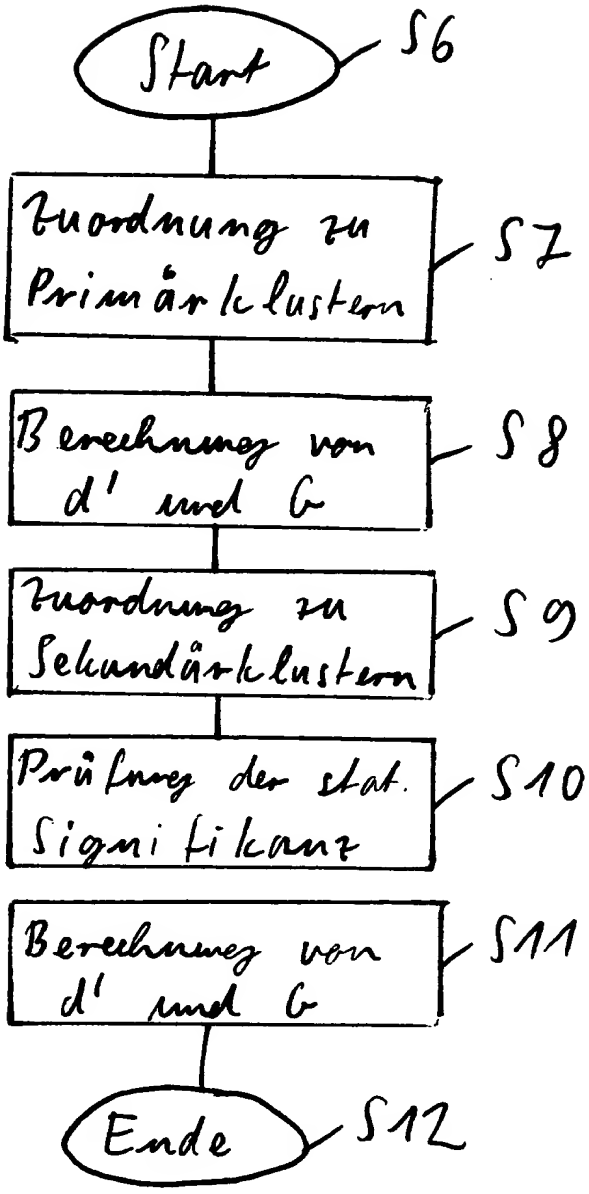

Fig. 2

2000 E 01067

2000 P 13225

SE2

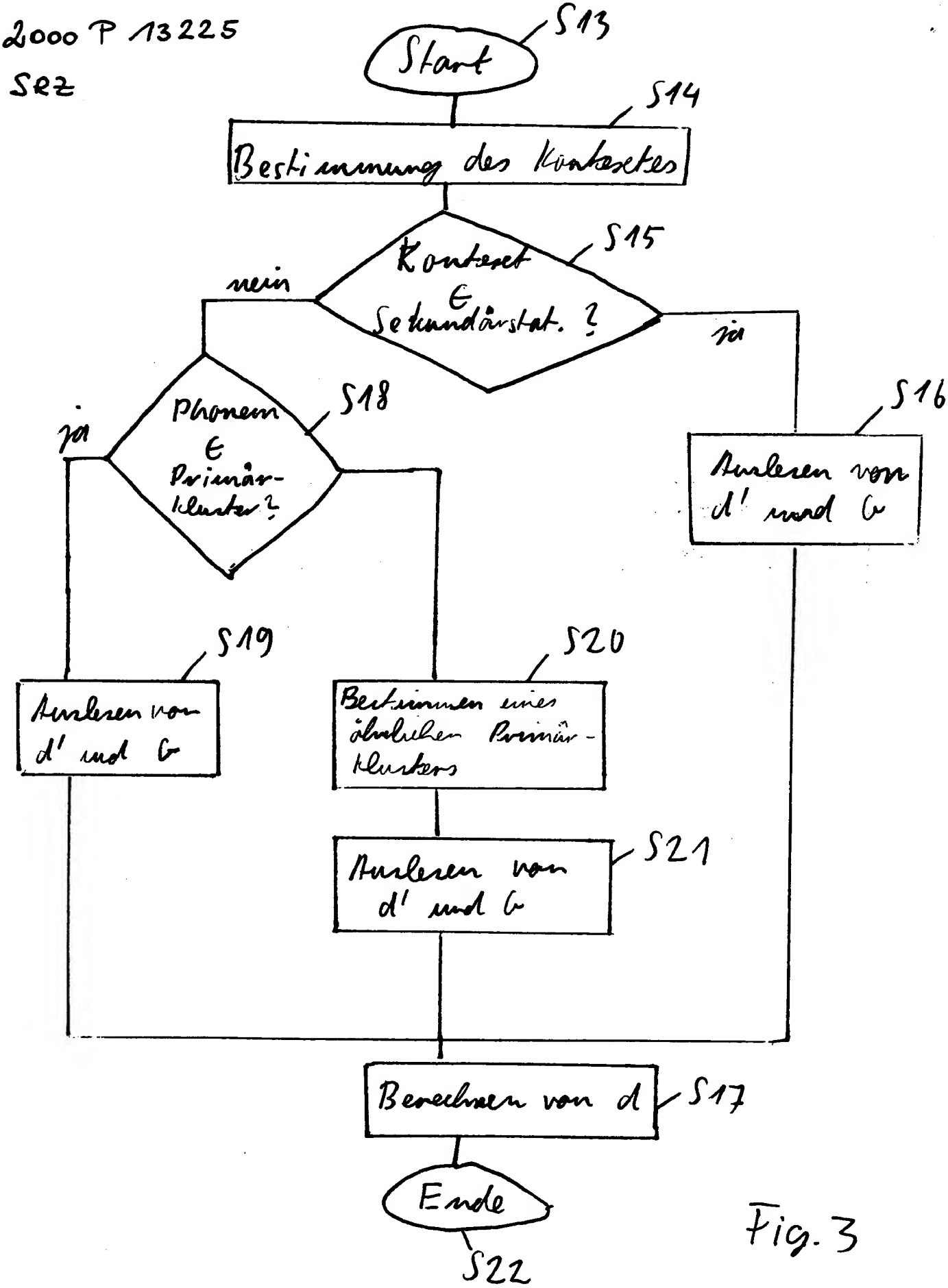

Fig. 3

2000 E 01067
2000 P 13225
SRZ

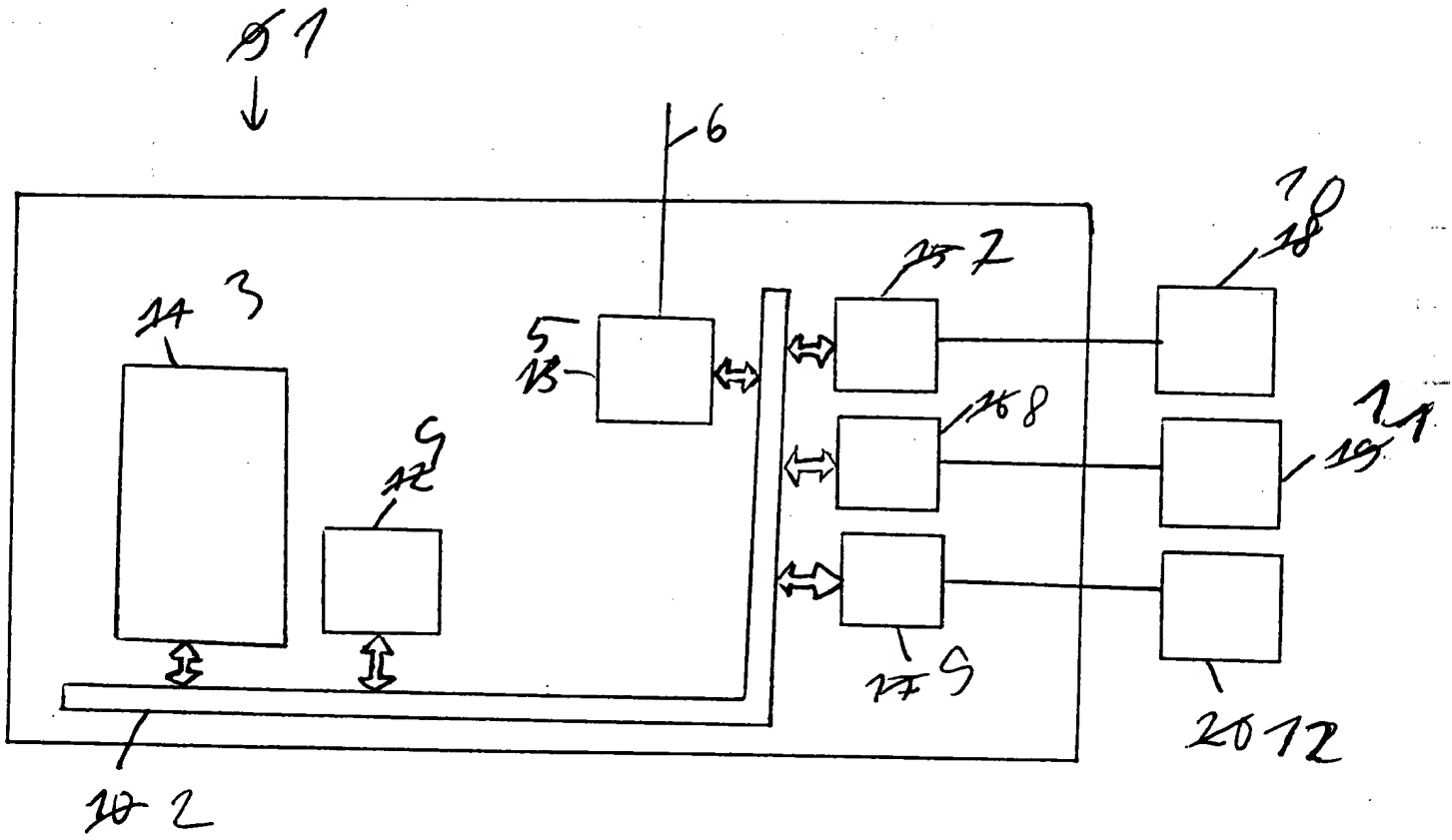

Fig. 4

THIS PAGE BLANK (USPTO)